Local Polynomial Estimation for Sensitivity Analysis on Models With Correlated Inputs

Sébastien da Veiga, Francois Wahl, Fabrice Gamboa

Local Polynomial Estimation for Sensitivity Analysis on Models With Correlated Inputs

Sebastien DA VEIGA[(1),(2)], Francois WAHL[(1)], Fabrice GAMBOA[(2)]
(1) : IFP-Lyon, France
(2) : Institut de Mathematiques, Toulouse, France

Abstract

Sensitivity indices when the inputs of a model are not independent are estimated by local polynomial techniques. Two original estimators based on local polynomial smoothers are proposed. Both have good theoretical properties which are exhibited and also illustrated through analytical examples. They are used to carry out a sensitivity analysis on a real case of a kinetic model with correlated parameters.

KEY WORDS: Nonparametric regression; Global sensitivity indices; Conditional moments estimation.

Achieving better knowledge of refining processes usually requires to build a kinetic model predicting the output components produced in a unit given the input components introduced (the "feed") and the operating conditions. Such a model is based on the choice of a reaction mechanism depending on various parameters (*e.g.* kinetic constants). But the complexity of the mechanism, the variability of the behavior of catalysts when they are used and the difficulty of observations and experiments imply that most often these parameters cannot be inferred from theoretical considerations and need to be estimated through practical experiments. This estimation procedure leads to consider them uncertain and this uncertainty spreads on the model predictions. This can be highly problematic in real situations. It is then essential to quantify this uncertainty and to study the influence of parameters variations on the model outputs through uncertainty and sensitivity analysis.

During the last decades much effort in mathematical analysis of physical processes has focused on modeling and reasoning with uncertainty and sensitivity. Model calibration and validation are examples where sensitivity and uncertainty analysis have become essential investigative scientific tools. Roughly speaking, uncertainty analysis refers to the inherent variations of a model (*e.g.* a modeled physical process) and is helpful in finding the relation between some variability or probability distribution on input parameters and the variability and probability distribution of outputs, while sensitivity analysis investigates the effects of varying a model input on the outputs and ascertains how much a model depends on each or some of its inputs.

Over the years several mathematical and computer-assisted methods have been developed to carry out global sensitivity analysis and the reader may refer to the book of Saltelli, Chan & Scott (2000) for a wide and thorough review. Amongst these methods a particular popular class is the one composed by "variance-based" methods which is detailed below. Let us consider a mathematical model given by

$$Y = \eta(\mathbf{X}) \tag{1}$$

where $\eta : \mathbb{R}^d \to \mathbb{R}$ is the modeling function, $Y \in \mathbb{R}$ represents the output or prediction of the model and $\mathbf{X} = (X_1, ..., X_d)$ is the d-dimensional real vector of the input factors or parameters. The vector of input parameters is treated as a random vector, which implies that the output is also a random variable. In variance-based methods, we are interested in explaining the variance $\mathrm{Var}(Y)$ through the variations of the X_i, $i = 1, ..., d$ and we decompose $\mathrm{Var}(Y)$ as follows :

$$\mathrm{Var}(Y) = \mathrm{Var}(\mathbb{E}(Y|X_i)) + \mathbb{E}(\mathrm{Var}(Y|X_i))$$

for all $i = 1, ..., d$ where $\mathbb{E}(Y|X_i)$ and $\mathrm{Var}(Y|X_i)$ are respectively the conditional expectation and variance of Y given X_i. The importance of X_i on the variance of Y is linked to how well $\mathbb{E}(Y|X_i)$ fits Y and can then be measured by the *first order sensitivity index*

$$S_i = \frac{\mathrm{Var}(\mathbb{E}(Y|X_i))}{\mathrm{Var}(Y)}$$

also called *correlation ratio*. We can also introduce sensitivity indices of higher orders to take into account input interactions. For example, the *second order sensitivity index* for X_i and X_j is

$$S_{ij} = \frac{\mathrm{Var}(\mathbb{E}(Y|X_i, X_j)) - \mathrm{Var}(\mathbb{E}(Y|X_i)) - \mathrm{Var}(\mathbb{E}(Y|X_j))}{\mathrm{Var}(Y)},$$

and so on for other orders, see Saltelli et al. (2000) for details.

In the case of independent inputs, two techniques, Sobol (Sobol' 1993) and FAST (Cukier, Fortuin, Shuler, Petschek & Schaibly 1973) are the most popular methods for estimating the S_i indices. Although powerful and computationally efficient, these methods rely on the assumption of independent

inputs which is hard to hold in many practical cases for kinetic models. Nevertheless, three original methods, originated by Ratto, Tarantola & Saltelli (2001), Jacques, Lavergne & Devictor (2004) and Oakley & O'Hagan (2004), try to deal with this problem. The first one sets out to calculate the sensitivity indices by using a replicated latin hypercube sampling, but this approach requires a large amount of model evaluations to reach an acceptable precision. The second one is based on the idea of building new sensitivity indices which generalize the original ones by taking into account block of correlations among the inputs. This method is however useless when many input factors are correlated. The last approach is that of Oakley & O'Hagan (2004) and rely upon the idea of approximating the function η in model (1) by a so-called 'kriging' response surface (Santner, Williams & Notz 2003) and of computing analytical expressions of the sensitivity indices based on the results of the kriging approximation. However appealing and accurate, these analytical expressions involve multidimensional integrals that are only tractable when the conditional densities of the input factors are known and easy to integrate. If this is not the case the multidimensional integrals must be approximated numerically, but at high computational cost. We then propose a new way of estimating sensitivity indices through an intermediate technique in the sense that it is based on a sample from the joint density of the inputs and the output like Ratto et al. (2001) but also on a nonparametric regression model like Oakley & O'Hagan (2004). This approach does not require as many model evaluations as Ratto et al. (2001) and does not require to approximate multidimensional integrals as Oakley & O'Hagan (2004) in the general case.

In this paper to deal with correlated inputs we consider a new method based on local polynomial approximations for conditional moments (see the work of Fan & Gijbels (1996) and Fan, Gijbels, Hu & Huang (1996) on conditional expectation and the papers of Fan & Yao (1998) and Ruppert, Wand, Holst & Hˆssjer (1997) on the conditional variance). Given the form of the sensitivity indices, local polynomial regression can be used to estimate them. This approach not only allows to compute a sensitivity index through an easy black-box procedure but also reaches a good precision.

The paper is organized as follows. In Section 1 we review the methods of Ratto et al. (2001), Jacques et al. (2004) and Oakley & O'Hagan (2004) and discuss their merits and drawbacks. In Section 2 we propose and study two new estimators for sensitivity indices relying on local polynomial methods. In Section 3 we present both analytical and practical examples. In Section 4 we finally give some conclusions and directions for future research.

1. MODELS WITH CORRELATED INPUTS

When the inputs are independent, Sobol showed that the sum of the sensitivity indices of all orders is equal to 1, due to an orthogonal decomposition of the function η (Sobol' 1993). Indeed sensitivity indices naturally arise from this functional ANOVA decomposition. Nevertheless, when the inputs are correlated, this property does not hold anymore because such a decomposition can not be done without taking into account the joint distribution of the inputs. If one decides to estimate sensitivity indices under the independence hypothesis although it does not hold, results and consequently interpretation can be higly misleading, see the first example of Section 3.1. But we can still consider the initial ANOVA decomposition and work with the original sensitivity indices without ignoring the correlation, and when quantifying the first order sensitivity index of a particular input factor a part of the sensitivity of all the other input factors correlated with it is also taken into account. Thus the same information is considered several times. Interpretation of sensitivity indices when the inputs are not independent becomes problematic. However, the input factors being independent or not, the first-order sensitivity index still points out which factor (if fixed) will mostly reduce the variance of the output. Thus, if the goal of the practitioner is to conduct a 'Factors Prioritisation' (Saltelli, Tarantola, Campolongo & Ratto 2004), *i.e.* identifying the factor that one should fix to achieve the greatest reduction in the uncertainty of the output, first-order sensitivity indices remain the measure of importance to study, see Saltelli et al. (2004). Considering that this goal is common for practitioners, being able to compute first-order sensitivity indices when the inputs are no longer independent is an interesting challenge.

Beyond this problem of interpretation, correlation also makes the computational methods FAST and Sobol unusable as they have been designed for the independent case. To get over these difficulties, it is first possible to build 'new' sensitivity indices that would generalize the original ones and match their properties, allowing interpretation. This is the idea of multidimensional sensitivity analysis of Jacques (Jacques et al. 2004) detailed in the next section. Secondly, Ratto et al. (2001) tried to continue on working with the original sensitivity indices and to compute them as described in Section 1.3, even if they do not give clues for interpretation. The authors generate replicated latin hypercube samples to approximate conditional densities. Finally, Oakley & O'Hagan (2004) suggest to approach the function η in model (1) by a kriging response surface which allows to get analytical expressions of sensitivity indices through multidimensional integrals.

1.1 Multidimensional Sensitivity Analysis

To define multidimensional sensitivity indices, Jacques et al. (2004) suggest to split $\mathbf{X}$ into p vectors $\mathbf{U}_j,\ j = 1, ..., p$, each of size k_j such that $\mathbf{U}_j$ is independent from $\mathbf{U}_l$ for $1 \leq j, l \leq p$, $j \neq l$:

$$\begin{aligned}\mathbf{X} = (X_1, ..., X_d) \quad = \quad & (\underbrace{X_1, ..., X_{k_1}}_{\mathbf{U}_1}, \underbrace{X_{k_1+1}, ..., X_{k_1+k_2}}_{\mathbf{U}_2}, ... \\ & ..., \underbrace{X_{k_1+k_2+...+k_{p-1}+1}, ..., X_{k_1+k_2+...+k_p}}_{\mathbf{U}_p})\end{aligned}$$

where $k_1 + k_2 + ... + k_p = d$. For example, if $\mathbf{X} = (X_1, X_2, X_3)$ where X_1 is independent of X_2 and X_3 but X_2 and X_3 are correlated, we set $\mathbf{U}_1 = X_1$ and $\mathbf{U}_2 = (X_2, X_3)$.

Thus they build *first order multidimensional sensitivity indices* using the $\mathbf{U}_j$ vectors :

$$S_j = \frac{\mathrm{Var}(\mathbb{E}(Y|\mathbf{U}_j))}{\mathrm{Var}(Y)} = \frac{\mathrm{Var}(\mathbb{E}(Y|X_{k_1+k_2+...+k_{j-1}+1}, ..., X_{k_1+k_2+...+k_j}))}{\mathrm{Var}(Y)}$$

for $j = 1, ..., p$. Remark that if the inputs are independent, these sensitivity indices have the same expression as in classical sensitivity analysis. Finally, it is possible to compute these indices by Monte-Carlo estimations.

Nevertheless, this method has a main drawback hard to overcome. If all the inputs are correlated, the $\mathbf{U}_j$ vectors cannot be defined (except the trivial case $\mathbf{U}_1 = \mathbf{X}$) and interpretation is not possible. The problem remains the same if too many inputs are dependent because this situation leads to consider very few multidimensional indices. Moreover, identifying a set of correlated variables $\mathbf{U}_j$ with high sensitivity index does not allow to point up whether this is due to one particular input of the set as we cannot differentiate among them. We will illustrate this phenomenon in the second example of Section 3.1.

1.2 Correlation-Ratios With Known Conditional Density Functions

The estimator introduced by Ratto et al. (2001) was first discussed in McKay (1996) and is based on samples from the conditional density functions of Y given X_i, $i = 1, ..., d$.

Let $(\mathbf{X}^j)_{j=1,...,n}$ be an i.i.d sample of size n from the distribution of the vector $\mathbf{X}$. $(X_i^j)_{j=1,...,n}$ is then an i.i.d. sample of size n from the distribution of the input factor X_i. For each realization X_i^j of this sample, let $(Y_i^{jk})_{k=1,...,r}$ be an i.i.d. sample of size r from the conditional density function of Y given $X_i = X_i^j$ and define the sample means

$$\overline{Y}_i^j = \frac{1}{r}\sum_{k=1}^{r} Y_i^{jk} \quad \overline{Y}_i = \frac{1}{n}\sum_{j=1}^{n} \overline{Y}_i^j.$$

Note that $\overline{Y}_i^j$ and $\frac{1}{r}\sum_{k=1}^{r}(Y_i^{jk} - \overline{Y}_i^j)^2$ respectively estimate the conditional expectation $\mathbb{E}(Y|X_i = X_i^j)$ and the conditional variance $\mathrm{Var}(Y|X_i = X_i^j)$, while $\overline{Y}_i$ estimates $\mathbb{E}(Y)$.

Using these moments estimators the numerator of the first order sensitivity index S_i, $\mathrm{Var}(\mathbb{E}(Y|X_i))$, can be estimated by the empirical estimator

$$\frac{1}{n}\sum_{j=1}^{n}(\overline{Y}_i^j - \overline{Y}_i)^2.$$

Similarly the denominator of S_i, $\mathrm{Var}(Y)$, is estimated by

$$\frac{1}{n}\sum_{j=1}^{n}\frac{1}{r}\sum_{k=1}^{r}(Y_i^{jk} - \overline{Y}_i)^2.$$

The estimator of the first order sensitivity index S_i of the input factor X_i, $i = 1, ..., d$ is then defined as

$$\hat{S}_i = \frac{SSB}{SST}$$

where

$$SSB = r\sum_{j=1}^{n}(\overline{Y}_i^j - \overline{Y}_i)^2$$

and

$$SST = \sum_{j=1}^{n}\sum_{k=1}^{r}(Y_i^{jk} - \overline{Y}_i)^2.$$

To compute these indices and to generate the samples needed, Ratto uses two different methods : pure Monte-Carlo sampling and a single replicated Latin HyperCube (r-LHS) sampling.

It is crucial to note, however, that these two methods require a huge amount of model evaluations to reach a good precision and can only be used for cases where model runs have very low computational cost.

1.3 Bayesian Sensitivity Analysis

The idea of Oakley & O'Hagan (2004) is to see the function $\eta(\cdot)$ as an unknown smooth function and to formulate a prior distribution for it. More precisely, it is modeled as the realization of a Gaussian stationary random field with given mean and covariance functions. Then, given a set of of values $y_i = \eta(\mathbf{x}_i)$, we can derive the posterior distribution of $\eta(\cdot)$ by classical Bayesian considerations. The prior distribution of $\eta(\mathbf{x})$ is a Gaussian stationary field :

$$\eta(\mathbf{x}) = \mathbf{h}(\mathbf{x})^t\beta + Z(\mathbf{x})$$

conditionally on β and σ^2, where $h(\cdot)$ is a vector of q known regression functions and $Z(\mathbf{X})$ is a Gaussian stationary random field with zero mean and covariance function $\sigma^2 c(\mathbf{x}, \mathbf{x}')$. The vector $h(\cdot)$ and the correlation function $c(\cdot, \cdot)$ are to be chosen in order to incorporate some information about how the output responds to the inputs and about the amount of smoothness we require on the output respectively. We refer the reader to Santner et al. (2003) and to Kennedy & O'Hagan (2001) for a detailed discussion on these choices. The second stage prior concerns the conjugate prior form for β and σ^2, which is chosen to be a normal inverse gamma distribution. Now assuming we observe a set $\mathbf{y}$ of n values of $y_i = \eta(\mathbf{x}_i)$, we can derive that the posterior distribution of $\eta(\cdot)$ given these data is a Student distribution, see Oakley & O'Hagan (2004) for details.
Using this posterior distribution, sensitivity indices can be computed analytically through multidimensional integrals involving functions of the observations and the conditional distributions of the input factors only. The main advantage of this Bayesian approach is that the model is only evaluated to calculate the quantities above, *i.e.* to 'fit' the response surface. Once this is done the estimation of sensitivity indices just involves the conditional distributions of the input factors. When the number of model runs is fixed, this method clearly reduces the standard errors of the estimated sensitivity indices obtained by Monte-Carlo methods such as Sobol (when the input factors are independent) and can still be used when the input factors are not independent.

However, the multidimensional integrals leading to the computation of the sensitivity indices, if not tractable analytically, need to be estimated. Let us describe more particularly one of the estimators proposed in Oakley & O'Hagan (2004). We

keep the authors notations and denote by $\mathbb{E}^*$ the expectations defined with respect to the posterior distribution of $\eta(\cdot)$. The numerator of the first-order sensitivity index of Y with respect to X_1 is estimated by

$$\mathbb{E}^*(\mathrm{Var}(\mathbb{E}(Y|X_1))) = \mathbb{E}^*(\mathbb{E}(\mathbb{E}(Y|X_1)^2)) - \mathbb{E}^*(\mathbb{E}(Y)^2)$$

and one of the quantities involved in the computation of $\mathbb{E}^*(\mathbb{E}(\mathbb{E}(Y|X_1)^2))$ is for example

$$\begin{aligned} U_1 \quad = \quad & \int_{\mathbb{R}^{d-1}} \int_{\mathbb{R}^{d-1}} \int_{\mathbb{R}} c(\mathbf{x}, \mathbf{x}^*)\ \mathrm{d}F_{-1|1}(\mathbf{x}_{-1}|x_1) \\ & \mathrm{d}F_{-1|1}(\mathbf{x'}_{-1}|x_1)\ \mathrm{d}F_1(x_1) \end{aligned}$$

where $F_{-1|1}$ is the marginal distribution of $\mathbf{X}_{-1}$ (subvector of $\mathbf{X}$ containing all elements except X_1) given X_1, F_1 is the marginal distribution of X_1 and $\mathbf{x}^*$ denotes the vector with elements made up of $\mathbf{x}_1$ and $\mathbf{x'}_{-1}$ in the same way as $\mathbf{x}$ is composed of $\mathbf{x}_1$ and $\mathbf{x}_{-1}$. If the conditional distribution $F_{-1|1}$ is not analytically known, we first need to estimate it with a sample of the joint distribution F. Many methods have been developed to do so, let us just mention for example kernel techniques. But in general in high dimension the data is very sparsely distributed and it is difficult to get an accurate approximation of conditional distributions since the so-called curse of dimensionality arises. For instance the best possible MSE rate with kernel techniques is $n^{-4/(4+d)}$ which becomes worse as d gets larger. Moreover, even if we could get a good approximation of $F_{-1|1}$, still remains the problem of evaluating the multidimensional integrals. Indeed the dimensionality of these integrals can reach $2d-1$ as in the expression of U_1 above. Since these integrals can not in general be separated into unidimensional integrals, approximating them with a sufficent accuracy is not an obvious mathematical problem. Deterministic schemes can not reasonably be considered, and with Monte-Carlo or quasi Monte-Carlo sampling (Owen 2005) thousands (or millions) of draws are required to get a reasonable accuracy.

With unknown densities, even if conceptually, sampling rather than analytical integration in the Oakley and O'Hagan approach seems reasonable, the results could be highly affected by the curse of dimensionality. Let us mention that Pr. O'Hagan has public domain software carrying out this analysis. However it does not yet allow to consider dependent inputs.

2. NEW ESTIMATION METHODOLOGY

Our approach is to estimate the conditional moments $\mathbb{E}(Y|X_i = X_i^j)$ and $\mathrm{Var}(Y|X_i = X_i^j)$ with an intermediate method between the one of Ratto et al. (2001) and Oakley & O'Hagan (2004). We first use a sample (X_i, Y_i) to estimate the conditional moments with nonparametric tools (provided they are smooth functions of the input factors). Then, we compute sensitivity indices by using another sample of the input factors only (and thus no more model runs are needed). While Oakley & O'Hagan (2004) approximate the function $\eta(\mathbf{X})$ in $\mathbb{R}^d$, we approximate it marginally, *i.e.* we approximate the conditional expectations $\mathbb{E}(\eta(\mathbf{X})|X_i)$ in $\mathbb{R}$. This approach allows to overcome the multidimensional integration problem of the Bayesian sensitivity analysis.

To simplify the notations, until Section 2.4 (X, Y) will stand for a bivariate random vector (*i.e.* X is unidimensional). As the variance may be decomposed as $\mathrm{Var}(Y) = \mathrm{Var}(\mathbb{E}(Y|X)) + \mathbb{E}(\mathrm{Var}(Y|X))$, the index we wish to estimate can be written

$$S = \frac{\mathrm{Var}(\mathbb{E}(Y|X))}{\mathrm{Var}(Y)} \quad \text{or} \quad S = 1 - \frac{\mathbb{E}(\mathrm{Var}(Y|X))}{\mathrm{Var}(Y)}. \qquad (2)$$

These expressions clearly give two ways of estimating S : the issue is to be able to estimate $\mathrm{Var}(\mathbb{E}(Y|X))$ or alternatively $\mathbb{E}(\mathrm{Var}(Y|X))$, obviously by estimating first the conditional moments $\mathbb{E}(Y|X = x)$ and $\mathrm{Var}(Y|X = x)$ $(x \in \mathbb{R})$. In both cases the denominator term $\mathrm{Var}(Y)$ can be easily estimated. To approximate the conditional moments, we propose to use local polynomial regression. This highly statistical efficient tool is easy to apprehend as it is close to the weighted least-squares approach in regression problems. Only basic results will be presented here, for a detailed picture of the subject the interested reader is referred to Fan & Gijbels (1996).

2.1 Formulation of the Estimators

Let $(X_i, Y_i)_{i=1,\ldots,n}$ be a two-dimensional i.i.d. sample of a real random vector (X, Y). Assuming that X and Y are square integrable we may write an heteroskedastic regression model of Y_i on X_i, exhibiting the conditional expectation and variance, as

$$Y_i = m(X_i) + \sigma(X_i)\epsilon_i, \qquad i = 1, \ldots, n$$

where $m(x) = \mathbb{E}(Y|X = x)$ and $\sigma^2(x) = \mathrm{Var}(Y|X = x)$ $(x \in \mathbb{R})$ are the conditional moments and the errors $\epsilon_1, \ldots, \epsilon_n$ are independent random variables satisfying $\mathbb{E}(\epsilon_i|X_i) = 0$ and $\mathrm{Var}(\epsilon_i|X_i) = 1$. Usually ϵ_i and X_i are assumed to be independent although this is not the case in our work. Note that results for correlated errors have been recently developed (Vilar-Fernández & Francisco-Fernández (2002) for the autoregressive case for example). Local polynomial fitting consists in approximating *locally* the regression function m by a p-th order polynomial

$$m(z) \approx \sum_{j=0}^{p} \beta_j (z - x)^j$$

for z in a neighborhood of x. This polynomial is then fitted to the observations (X_i, Y_i) by solving the weighted least-squares problem

$$\min_{\beta} \sum_{i=1}^{n} \Big(Y_i - \sum_{j=0}^{p} \beta_j (X_i - x)^j \Big)^2 K_1\Big(\frac{X_i - x}{h_1}\Big) \qquad (3)$$

where $K_1(.)$ denotes a *kernel* function and h_1 is a *smoothing* parameter (or *bandwidth*). In this case, if $\hat{\beta}(x) = (\hat{\beta}_0(x), ..., \hat{\beta}_p(x))^T$ denotes the minimizer of (3) we have

$$\hat{m}(x) = \hat{\beta}_0(x),$$

while the ν-th derivative of $m(x)$ is estimated via the relation

$$\hat{\beta}_\nu(x) = \frac{\hat{m}^{(\nu)}(x)}{\nu!},$$

see Fan & Gijbels (1996) for more details. As it will be discussed later, the smoothing parameter h_1 is chosen to balance bias and variance of the estimator. Finally, remark that the particular case $p = 0$ (constant fit) leads to the well-known *Nadaraya-Watson* estimator $\hat{m}_{NW}(x)$ of the conditional expectation, given explicitly by

$$\hat{m}_{NW}(x) = \frac{\sum_{i=1}^{n} Y_i K\Big(\frac{X_i - x}{h}\Big)}{\sum_{i=1}^{n} K\Big(\frac{X_i - x}{h}\Big)},$$

see Wand & Jones (1994).

Estimation of the conditional variance is less straightforward. If the regression function m was known, the problem of estimating $\sigma^2(.)$ would be regarded as a local polynomial regression of r_i^2 on X_i with $r_i^2 = (Y_i - m(X_i))^2$, as $\mathbb{E}(r^2|X = x) = \sigma^2(x)$ with $r^2 = (Y - m(X))^2$. But in practice, m is unknown. A natural approach is to substitute $m(.)$ by its estimate $\hat{m}(.)$ defined as above and to get the *the residual-based estimator* $\hat{\sigma}^2(x)$ by solving as previously the weighted least-squares problem

$$\min_{\gamma} \sum_{i=1}^{n} \Big(\hat{r}_i^2 - \sum_{j=1}^{q} \gamma_j (X_i - x)^j\Big)^2 K_2\Big(\frac{X_i - x}{h_2}\Big) \qquad (4)$$

where $\hat{r}_i^2 = (Y_i - \hat{m}(X_i))^2$, $K_2(.)$ is a kernel and h_2 a smoothing parameter. Note that the kernel $K_2(.)$ is not necessarily chosen to be equal to the kernel $K_1(.)$. Then

$$\hat{\sigma}^2(x) = \hat{\gamma}_0(x)$$

where $\hat{\gamma}(x) = (\hat{\gamma}_0(x), ..., \hat{\gamma}_q(x))$ is the minimizer of (4). As previously, the smoothing parameter h_2 has to be chosen to balance bias and variance of the estimator, see Fan & Yao (1998).

Going back over the equalities in (2), the last step is to estimate the quantities $\mathrm{Var}(\mathbb{E}(Y|X))$ and $\mathbb{E}(\mathrm{Var}(Y|X))$ by using the local polynomial estimators for the conditional moments defined right above. To do this let us assume we have another i.i.d. sample $(\tilde{X}_j)_{j=1,...,n'}$ with same distribution as X. If the functions $m(.)$ and $\sigma^2(.)$ were known, we could estimate $\mathrm{Var}(\mathbb{E}(Y|X)) = \mathrm{Var}(m(X))$ and $\mathbb{E}(\mathrm{Var}(Y|X)) = \mathbb{E}(\sigma^2(X))$ with the classical empirical moments

$$\frac{1}{n'-1} \sum_{j=1}^{n'} \Big(m(\tilde{X}_j) - \bar{m}\Big)^2 \quad \text{and} \quad \frac{1}{n'} \sum_{j=1}^{n'} \sigma^2(\tilde{X}_j)$$

where $\bar{m} = \dfrac{1}{n'} \sum_{j=1}^{n'} m(\tilde{X}_j)$. As $m(.)$ and $\sigma^2(.)$ are unknown, the main idea is to replace them by their local polynomial estimators which leads to consider

$$\hat{T}_1 = \frac{1}{n'-1} \sum_{j=1}^{n'} \Big(\hat{m}(\tilde{X}_j) - \hat{\bar{m}}\Big)^2 \quad \text{and} \quad \hat{T}_2 = \frac{1}{n'} \sum_{j=1}^{n'} \hat{\sigma}^2(\tilde{X}_j)$$

where $\hat{\bar{m}} = \dfrac{1}{n'} \sum_{j=1}^{n'} \hat{m}(\tilde{X}_j)$ and $\hat{m}(.)$ and $\hat{\sigma}^2(.)$ are the local polynomial estimators of $m(.)$ and $\sigma^2(.)$ introduced above. It is important to note that we need two samples, the first one $(X_i, Y_i)_{i=1,...,n}$ to compute $\hat{m}(.)$ and $\hat{\sigma}^2(.)$ and the second one $(\tilde{X}_j)_{j=1,...,n'}$ to finally compute the empirical estimators $\hat{T}_1$ and $\hat{T}_2$.

2.2 Bandwidth and Orders Selection

The selection of the smoothing parameters h_1 and h_2 and to a lesser extent of the polynomials orders p and q can be crucial to get the least mean squared error (MSE) of the estimators $\hat{T}_1$ and $\hat{T}_2$. Classically the MSE consists of a bias term plus a variance term and so is minimized by finding a compromise between bias and variance.

Concerning this choice, the reader is referred to Fan et al. (1996), Fan & Yao (1998) or Ruppert (1997). Most of the methods suggested by these authors rely upon asymptotic arguments and their efficiency for finite sample cases is not clear. In practice cross-validation methods can be used for the finite sample case (Jones, Marron & Sheather 1996), but in the examples of Section 3 we will use the empirical-bias bandwidth selector (EBBS) of Ruppert which appears to be efficient on simulated data. EBBS is based on estimating the MSE empirically and not with an asymptotic expression. The choice of the polynomials orders is more subjective. Concerning the estimation of the conditional expectation, Fan & Gijbels (1996) recommend to use a $\nu+1$ or $\nu+3$th-order polynomial to estimate the νth-derivative of $m(x)$, following theoretical considerations on the asymptotic bias of $\hat{m}(x)$ on the boundary. We would then be lead to take $p = 1$ or $p = 3$ to estimate the 0th-derivative $m(x)$. But Ruppert, Wand & Carroll (2003) suggest that this conclusion should be balanced by simulation studies and stress that $p = 2$ often outperforms $p = 1$ and $p = 3$. The only common conclusion is that local linear regression ($p = 1$) is usually superior to kernel regression (Nadaraya-Watson estimator obtained with $p = 0$). This is the reason why we will only consider and study local linear regression for $m(x)$ in the next theoretical and practical sections. The choice is still difficult when estimating the conditional variance as we have to choose p and q simultaneously. One more time, the authors are not unanimous : Fan & Yao (1998) recommend the case $p = 1, q = 1$ whereas Ruppert et al. (1997) suggest $p = 2, q = 1$ or $p = 3, q = 1$. However on the simulations we have carried out, the choice of $p = 1, q = 1$ is adequate and satisfactory in terms of precision. This is the reason why we have decided to consider only the case $p = 1, q = 1$ for both theoretical and practical results.

2.3 Theoretical Properties of the Estimators

The properties of $\hat{T}_1$ and $\hat{T}_2$ strongly depend on the asymptotic results on the bias and variance of the local linear estimators $\hat{m}(.)$ and $\hat{\sigma}^2(.)$. We only give here two main results, all assumptions $(A_0, ..., A_4, B_0, ..., B_4, C_0)$ and proofs are given in appendix for readability. $\mathbb{E}_{\mathbb{X}}$ and $\mathrm{Var}_{\mathbb{X}}$ stand for the conditional expectation and variance given the predictors $\mathbb{X} = (X_1, ..., X_n)$. The expression $o_P(\varphi(h))$ is equal to $\varphi(h) o_P(1)$ for a given function φ. Here $o_P(1)$ is the standard notation for a sequence of random variables that converges to zero in probability.

Theorem 1 *Under assumptions (A0)-(A4) and (C0), the estimator $\hat{T}_1$ is asymptotically unbiased. More precisely*

$$\mathbb{E}_{\mathbb{X}}(\hat{T}_1) = Var(\mathbb{E}(Y|X)) + M_1 h_1^2 + \frac{M_2}{nh_1} + o_P(h_1^2).$$

where M_1 and M_2 are constants given in appendix.

Remark 1. It would be interesting to calculate the variance of this estimator, but it would require the expressions of the third and fourth moments of the local linear estimator $\hat{m}(.)$ (see the appendix). This is not an obvious problem and to the best of our knowledge it has not been addressed in the literature. It is beyond the scope of the present paper but it is an interesting problem for future research. Nevertheless, the variance can be estimated on practical cases through bootstrap methods for example (Efron & Tibshirani 1994).

Theorem 2 *Under assumptions (B0)-(B4) and (C0), the estimator $\hat{T}_2$ is consistent. More precisely*

$$\mathbb{E}_{\mathbb{X}}(\hat{T}_2) = \mathbb{E}(\, Var(Y|X)) + V_1 h_2^2 + o_P(h_1^2 + h_2^2)$$

and

$$\begin{aligned} Var_{\mathbb{X}}(\hat{T}_2) &= \frac{1}{n'}\left\{\mathbb{E}(\, Var(Y|X)^2) + V_2 h_2^2 + V_3 h_1^2 + \frac{V_4}{nh_2}\right. \\ &\quad \left. + o_P\left(h_1^2 + h_2^2 + \frac{1}{\sqrt{nh_2}}\right)\right\} \end{aligned}$$

where V_1, V_2, V_3 and V_4 are constants given in appendix.

2.4 Application to Sensitivity Analysis

Let us come back to the model (1), where $\mathbf{X}$ is multidimensional. The goal is to get an estimate of S_i for $i = 1, \ldots, d$ by using one of the two estimators $\hat{T}_1$ and $\hat{T}_2$. We need two samples to compute each of them, *i.e.* a sample $(X_i^k, Y^k)_{k=1,\ldots,n}$ to estimate $\hat{m}(.)$ and $\hat{\sigma}^2(.)$ and a sample $(\tilde{X}_i^l)_{l=1,\ldots,n'}$ to get $\hat{T}_1$ and $\hat{T}_2$ where $(X_i^k)_{k=1,\ldots,n}$ and $(\tilde{X}_i^l)_{l=1,\ldots,n'}$ are samples from the joint distribution of the d-dimensional input factors $\mathbf{X} = (X_i)_{i=1,\ldots,d}$ and $(Y^k)_{k=1,\ldots,n}$ a sample of the output Y. Note that the model is run just for the first sample and not for the second one. Three situations can arise :

1. Sampling from the joint distribution of $\mathbf{X}$ has low computational cost and running the model to compute $(Y^k)_{k=1,\ldots,n}$ is cheap. This is the ideal situation. Indeed in this case the two samples $(X_i^k, Y^k)_{k=1,\ldots,n}$ and $(\tilde{X}_i^l)_{l=1,\ldots,n'}$ can be generated independently and be as large as required ;

2. Sampling from the joint distribution of $\mathbf{X}$ has low computational cost but model evaluations have not. In this case (also pointed out by Oakley & O'Hagan (2004)) a moderate-sized sample $(X_i^k, Y^k)_{k=1,\ldots,n}$ is used in order to fit the conditional moments. However to compute $\hat{T}_1$ and $\hat{T}_2$ we can then use a sample $(\tilde{X}_i^l)_{l=1,\ldots,n'}$ of large size ;

3. Sampling from the joint distribution of $\mathbf{X}$ has high computational cost. This case can arise in practice for example when the input factors are obtained through a procedure based on experimental data and optimization routines. We then have an initial sample $(\mathbf{X}^j)_{j=1,\ldots,N}$ of limited size N that we wish to use for the two steps of the estimation. The first idea is to split it and to use the first part to get the sample $(X_i^k, Y^k)_{k=1,\ldots,n}$ and the second one to get $(\tilde{X}_i^l)_{l=1,\ldots,n'}$. The drawback of this method clearly arises if N is very small. Another way to tackle the problem is to use the well-known leave-one-out idea procedure which gives better approximation than data splitting.
   As suggested by the Associate Editor another possible method could be to use the sample of size N to estimate the conditional moments and to estimate also the marginal densities of each input using for instance a nonparametric density estimator. One could then use these density estimates to get the sample $(\tilde{X}_i^l)_{l=1,\ldots,n'}$. The clear disadvantage of this procedure is that it may bias the final estimators. Some simulation runs not reported here for lack of space show that such a procedure leads to less efficient estimates probably due to the large bias produced by nonparametric methods.

The last situation obviously leads to the less accurate approximations of first-order sensitivity indices. However in general, litterature and results on sensitivity analysis assume that, if not analytically known, the joint distribution of the input factors can at least be generated at low computational cost. This is the reason why we will only describe here the procedure for estimating first-order sensitivity indices in case 1 or 2. We now assume that we have two samples $(X_i^k)_{k=1,\ldots,n}$ and $(\tilde{X}_i^l)_{l=1,\ldots,n'}$ obtained by one of the methods described right above.

The estimation procedure for $S_i = \dfrac{\mathrm{Var}(\mathbb{E}(Y|X_i))}{\mathrm{Var}(Y)}$ is the following :

Step 1 : Compute the output sample $(Y^k)_{k=1,\ldots,n}$ by running the model at $(\mathbf{X}^k)_{k=1,\ldots,n}$

Step 2 : Compute $\hat{\sigma}_Y^2$, the classical unbiased estimator of the variance $\mathrm{Var}(Y)$

$$\hat{\sigma}_Y^2 = \frac{1}{n-1}\sum_{k=1}^{n}\left(Y^k - \bar{Y}\right)^2$$

Step 3 : Use the sample $(X_i^k, Y^k)_{k=1,\ldots,n}$ to obtain $\hat{m}(\tilde{X}_i^l)$ for $l = 1, \ldots, n'$ and $\hat{m}(X_i^k)$ for $k = 1, \ldots, n$ using the smoothing parameter h_1 given by EBBS

Step 4 : Compute squared residuals $\hat{r}_k = (Y^k - \hat{m}(X_i^k))^2$ for $k = 1, \ldots, n$ and apply the smoothing parameter h_2 obtained by EBBS to compute $\hat{\sigma}^2(\tilde{X}_i^l)$ for $l = 1, \ldots, n'$

Step 5 : Compute $\hat{T}_1$ with $\hat{m}(\tilde{X}_i^l)$ for $l = 1, \ldots, n'$ from Step 3 and compute $\hat{T}_2$ with $\hat{\sigma}^2(\tilde{X}_i^l)$ for $l = 1, \ldots, n'$ from Step 4

Step 6 : The estimates of S_i are then

$$\hat{S}_i^{(1)} = \frac{\hat{T}_1}{\hat{\sigma}_Y^2} \quad \text{and} \quad \hat{S}_i^{(2)} = 1 - \frac{\hat{T}_2}{\hat{\sigma}_Y^2}. \tag{5}$$

To obtain all the first-order sensitivity indices, repeat the procedure from Step 3 to Step 6 for $i = 1, ..., d$.

Remark 2. Given the theoretical properties of $\hat{T}_1$ and $\hat{T}_2$ and more precisely their non-parametric convergence rate, we can also expect a nonparametric convergence rate for $\hat{S}^{(1)}$ and $\hat{S}^{(2)}$.

Remark 3. In practice, our simulations show that n of the order of 100 and n' around 2000 are enough for accurate estimation of the sensitivity indices.

3. EXAMPLES

In all the following examples we use the two estimators $\hat{S}^{(1)}$ and $\hat{S}^{(2)}$ defined in (5). As mentioned in Section 2.2, the conditional expectation is estimated here with local linear regression ($p = 1$) and the conditional variance with $p = 1$ and $q = 1$, the bandwidths being selected by the estimated-bias method of Ruppert (1997).

3.1 Analytical Examples

In this section, we carry out two different comparisons in order to study our two estimators from a numerical point of view. The first model has been chosen to underline their precision in correlated cases when FAST and Sobol methods are no longer efficient and when Jacques' approach for multidimensional sensitivity analysis is limited. We also show how interpretation with sensitivity indices obtained by neglecting correlation can be false. The second one is an example illustrating the performance of our estimators with respect to the method of Oakley and O'Hagan in a two-dimensional setting.

In the first analytical example, we study the model

$$Y = X_1 + X_2 + X_3$$

where (X_1, X_2, X_3) is a three-dimensional normal vector with mean $\mathbf{0}$ and covariance matrix

$$\Gamma = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & \rho\sigma \\ 0 & \rho\sigma & \sigma^2 \end{bmatrix}$$

where ρ is the correlation of X_2 and X_3 and $\sigma > 0$ is the standard deviation of X_3. The first order sensitivity indices can be evaluated analytically :

$$S_1 = \frac{1}{2 + \sigma^2 + 2\rho\sigma}$$

$$S_2 = \frac{(1 + \rho\sigma)^2}{2 + \sigma^2 + 2\rho\sigma}$$

$$S_3 = \frac{(\sigma + \rho)^2}{2 + \sigma^2 + 2\rho\sigma}$$

The first crucial remark to be done in this case is that we must take into account correlations to estimate sensitivity indices if we want a serious investigation of this model. Indeed, let us consider the case where $\sigma = 1.2$ and $\rho = -0.8$. We then have

$$S_1 = 0.6579, \quad S_2 = 0.0011, \quad S_3 = 0.1053,$$

indicating that X_1 should be the input to be fixed to reach the higher variance reduction on Y. But if one neglects the correlation, by computing for instance these indices with the FAST method, *i.e.* working with a three-dimensional normal vector with mean $\mathbf{0}$ and covariance matrix I instead of Γ, one would estimate

$$S_1^0 = 0.2907, \quad S_2^0 = 0.2907, \quad S_3^0 = 0.4186$$

where S^0 stands for the sensitivity indices when $\rho = 0$. These results indicate that X_3 should be fixed to mostly reduce the variance of Y, which is absolutely wrong as the calculations above have shown. This simple example highlights the danger of neglecting the correlations between the inputs and the importance to take them into consideration when computing sensitivity indices.

Otherwise, applying Jacques' idea to X_1 and the couple (X_2, X_3), we also get the expression of the first order multidimensional sensitivity index

$$S_{\{2,3\}} = \frac{1 + \sigma^2 + 2\rho\sigma}{2 + \sigma^2 + 2\rho\sigma}$$

Choosing $\rho = -0.2$ and $\sigma = 0.4$, we have

$$S_1 = S_{\{2,3\}} = 0.5, \quad S_2 = 0.4232, \quad S_3 = 0.02$$

If we interpret these indices as suggested by Jacques' multidimensional sensitivity analysis, the only conclusion we can give is that the couple (X_2, X_3) has the same importance as X_1. Indeed $S_{\{2,3\}} = S_1$. But actually the high value of $S_{\{2,3\}}$ comes from X_2 as shown by the exact calculations above, which implies that the information on $S_{\{2,3\}}$ alone is not sufficient. But with our method, we can estimate all the first order sensitivity indices :

$$\hat{S}_1^{(1)} = 0.4895, \quad \hat{S}_2^{(1)} = 0.4250, \quad \hat{S}_3^{(1)} = 0.0234$$

$$\hat{S}_1^{(2)} = 0.5081, \quad \hat{S}_2^{(2)} = 0.4368, \quad \hat{S}_3^{(2)} = 0.0361$$

for an average upon 100 simulations with $n = 50$ and $n' = 1000$. We display in Figure 1 the boxplots corresponding to the distribution of the sensitivity indices on these 100 simulations with the estimator $\hat{T}_2$. Because of the mathematical complexity mentioned before for the computation of the variance of $\hat{T}_1$, we are not able to recommend one estimator over the other one from a theoretical point of view. But in practice, we have observed that the variance of $\hat{T}_2$ is at least comparable to the variance of $\hat{T}_1$, and sometimes lower. Nervertheless, the computation of $\hat{T}_2$ is more difficult as illustrated in Section 2.4.

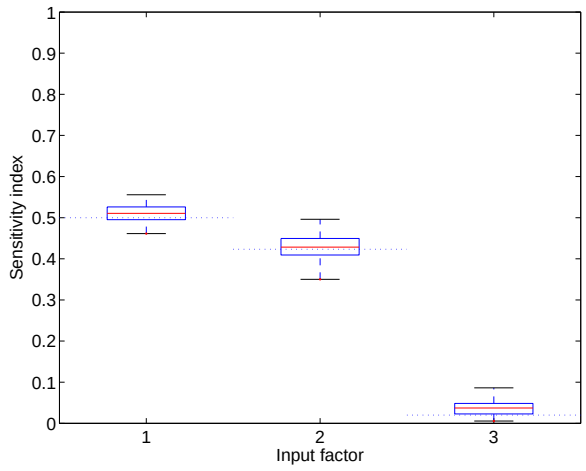

Figure 1. Boxplot of the estimated sensitivity indices ($\hat{S}^{(2)}$) for the three-factor additive model, 100 simulations. Dot lines are the true values.

Computing S_2 and S_3 with our method, even if both of them take into account correlations, allows to confirm the expected result : all the variability comes from X_2, and not from X_3. This simple example then brings out the limitation of the multidimensional approach.

In the second analytical example we consider the model

$$\begin{aligned} Y &= 0.2\exp(X_1-3)+2.2|X_2|+1.3X_2^6-2X_2^2 \\ &\quad -0.5X_2^4-0.5X_1^4+2.5X_1^2+0.7X_1^3 \\ &\quad +\frac{3}{(8X_1-2)^2+(5X_2-3)^2+1}+\sin(5X_1)\cos(3X_1^2) \end{aligned}$$

where X_1 and X_2 are independent random variables uniformly distributed on $[-1,1]$. Such a model is routinely used at Institut Francais du Petrole to compare different response surface methodologies as it presents a peak and valleys. The function is plotted in Figure 2.

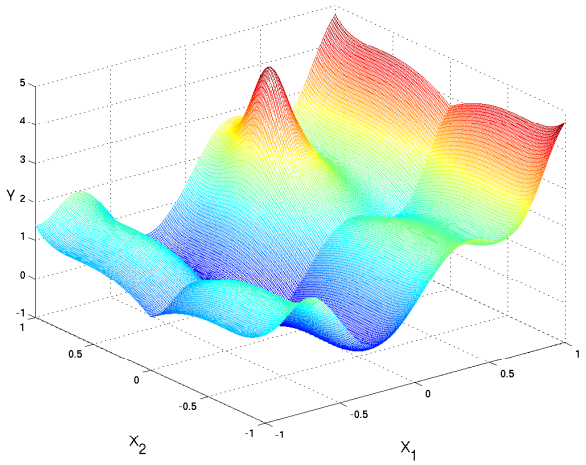

Figure 2. Function proposed in model 2 .

In this case, the sensitivity indices are

$$S_1 = 0.9375 \quad \text{and} \quad S_2 = 0.0625$$

We considered a 6×6 regular grid on $[0,1]^2$ and used it to estimate the posterior distribution in the method of Oakley and O'Hagan and to estimate the conditional moments in our method. Then, we calculated analytically the multidimensional integrals in the Bayesian approach while using a sample of size 5000 to compute $\hat{S}_i^{(2)}$ for $i = 1, 2$. The Bayesian approach leads to

$$\hat{S}_1 = 0.9038 \quad \text{and} \quad \hat{S}_2 = 0.0961$$

while the local polynomial technique gives

$$\hat{S}_1^{(2)} = 0.9127 \quad \text{and} \quad \hat{S}_2^{(2)} = 0.0452.$$

We can see on this example that the results obtained with both methods are comparable. However on this simple case the multidimensional integrals were analytically computed, which could not be the case in a non-independent setting. If not, a numerical integration, if feasible, would lead to less accurate approximations as discussed in Section 1.3.

3.2 Practical Example from Chemical Field : Isomerization of the Normal Butane

The isomerization of the normal butane, *i.e.* molecules with four carbon atoms, is a chemical process aiming at transforming normal butane (nC4) into iso-butane (iC4) in order to obtain a higher octane number, favored by iC. A simplified reaction mechanism has been used :

$$\text{nC4} \longleftrightarrow \text{iC4} \qquad (1)$$

$$2\ \text{iC4} \longrightarrow \text{C3+C5} \qquad (3)$$

$$\text{nC4} + \text{iC4} \longrightarrow \text{C3+C5} \qquad (4)$$

where reaction (1) is the main reversible reaction converting the normal butane into iso-butane. Reactions (3) and (4) are secondary and irreversible reactions which produce propane (C3) and a lump of normal and iso-pentane (C5), paraffins with three and five carbon atoms. The model linked to this process can be written as

$$\mathbf{Y} = \eta(\mathbf{c}, \boldsymbol{\theta})$$

where

- $\mathbf{Y}$ is the 3-dimensional result vector (mole fractions of the components nC4, iC4, C3 and C5 ; note that their sum is 1),

- $\mathbf{c}$ is the vector containing the operating conditions (pressure, temperature,...) and the mole fraction of the input components (nC4 and iC4, this is called the *feed*),

- $\boldsymbol{\theta} = (\theta^i)_{i=1,\dots,8}$ is the 8-dimensional random vector of the parameters of the reactions (pre-exponential factors, activation energies, adsorption constants,...),

- f is the function modeling the chemical reactor in which the reaction takes place. It is evaluated through the resolution of an ordinary differential equations system which can not be analytically solved and is calculated numerically.

The first step here is to get the distribution of $\boldsymbol{\theta}$ which is unknown. However, it is possible to use the experience and the

knowledge of chemical engineers to suggest a reasonable approximation of this distribution. Classically, we assume that $\boldsymbol{\theta}$ has a multivariate Gaussian distribution with mean zero (once the parameters are centered). Concerning the correlation matrix, it is built with experts and with the help of bootstrap simulations and is given by :

$$\Gamma = \begin{bmatrix} 1 & 0.43 & 0.09 & 0.29 & 0.55 & 0.66 & 0.10 & -0.01 \\ 0.43 & 1 & -0.54 & 0.11 & 0.37 & 0.25 & 0.51 & -0.48 \\ 0.09 & -0.54 & 1 & -0.02 & 0.20 & 0.02 & -0.40 & 0.73 \\ 0.29 & 0.11 & -0.02 & 1 & -0.41 & -0.07 & -0.22 & 0.01 \\ 0.55 & 0.37 & 0.20 & -0.41 & 1 & 0.43 & 0.31 & 0 \\ 0.66 & 0.25 & 0.02 & -0.07 & 0.43 & 1 & 0.17 & -0.11 \\ 0.10 & 0.51 & -0.40 & -0.22 & 0.31 & 0.17 & 1 & -0.61 \\ -0.01 & -0.48 & 0.73 & 0.01 & 0 & -0.11 & -0.61 & 1 \end{bmatrix}$$

In order to compute sensitivity indices, we generate a sample of size $n = 5000$ from this distribution.
Here we wish to estimate, for a given operating conditions and feed vector c, the sensitivity indices of the outputs with respect to the input factors in $\boldsymbol{\theta}$, *i.e.*

$$S_i^j = \frac{\mathrm{Var}(\mathbb{E}(Y_j|\theta^i))}{\mathrm{Var}(Y_j)}$$

for $j = 1, ..., 3$ and $i = 1, ..., 8$. Actually, our goal is to identify on which factor we should make the effort of reducing the uncertainty, by carrying out new experiments. This factor should be chosen in order to reduce as much as possible the uncertainty of the outputs.
We consider two particular vectors c_1 and c_2 containing the same operating conditions but a different feed (c_1 : nC4=1 and iC4=0, c_2 : iC4=1 and nC4=0). We have drawn for each vector c_i, $i = 1, 2$ a sample of size n from $\mathbf{Y}$ by Monte-Carlo simulations, *i.e.* by computing $\mathbf{Y}_j = \eta(c_i, \boldsymbol{\theta}_j)$ for $j = 1, ..., n$. Thus we have a sample from $(\mathbf{Y}, \boldsymbol{\theta})$ for each particular c_1 and c_2. For instance, the estimates of the sensitivity indices of the third output C3+C5 with the $\hat{T}_1$ estimator are given in Figure 4. Filled bars correspond to c_1 and empty bars to c_2.

Note that the estimates given by the $\hat{T}_2$ estimator are similar. These results highlight the behavior of the C3+C5 output when the feed changes. Indeed when we only use nC4 in the feed (c_1) the production of C3+C5 is mainly linked to the production of iC4 by reaction (1). This is confirmed by the importance of parameters 1 and 6 in Figure 4 which are the parameters involved in reaction (1). When the feed only contains iC4 (c_2), the first reaction is no longer dominating for the production of C3+C5, now mainly linked to reaction (3). Parameters 4 and 2 that are the most important in Figure 4 for c_2 are connected to reaction (3). We can thus conclude that the results confirm the expected behavior of the C3+C5 output.

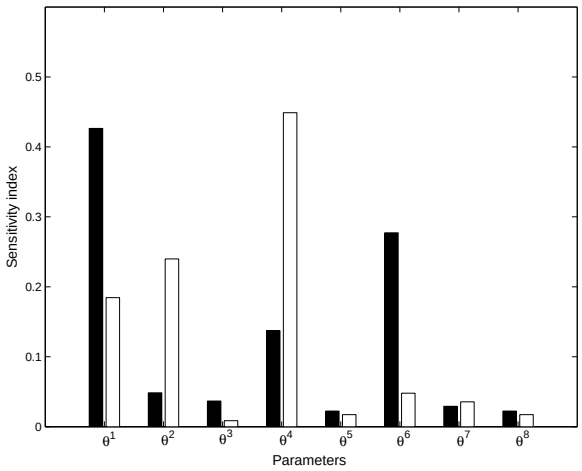

Figure 4. Sensitivity indices of the C3+C5 output in the isomerization model for the particular conditions c_1 (filled bars) and c_2 (empty bars).

We could obviously study the sensitivity indices for the other outputs, and for other operating conditions. Such a study has been carried out and showed that the most influent parameters depend on the operating conditions and the feed. But it also underlined that each parameter of the model has an influence on at least one output for at least one operating condition. In this case these sensitivity indices estimates enlighten the fact that all the parameters are potentially important. A discussion with chemical engineers would then be necessary in order to identify which outputs are most critical for their goals (controlling for instance the first output iC4 which is strongly linked to the octane rate) and would thus help us to choose which input parameters deserve most attention.

4. DISCUSSION AND CONCLUSION

The estimation method proposed in this paper is an efficient way to carry out sensitivity analysis by computing first order sensitivity indices when the inputs are not independent. The use of local polynomial estimators is the key point of the estimation procedure. It guarantees some interesting theoretical properties and ensures good qualities to the estimators we have introduced. Beyond these theoretical results, practical examples also show a good precision for a rather low computation time. Obviously, higher precision requires higher calculation time and the user has the possibility to adapt the estimators, by fixing some hyper-parameter values such as polynomials orders.

The main advantage of our estimators is obviously that they only make the assumption that the marginals are smooth and then require less model runs than classical sampling methods. Comparing with the Bayesian approach of Oakley & O'Hagan (2004), our method has the same philosophy as it uses model runs to fit a response surface under smoothness assumptions, but we avoid its numerical integration issue in high dimension. Moreover our approach is appealing for practioners in the sense that they can see it as a black-box routine, as each step of the procedure is data-driven once the user has given the two samples needed for the estimation.

Finally, we think that a practitioner willing to carry out a sensitivity analysis should combine different approach to get the most accurate result, for example computing the indices with the method we introduce her and the one of Oakley and O'Hagan. Indeed these two methods are not concurrent but complementary.

Future work will also be based on building multi-outputs sensitivity indices through multivariate nonparametric regression techniques.

ACKNOWLEDGMENTS

We thank Professor Anestis Antoniadis for very helpful discussions. We also thank the referees and associate editor for very useful suggestions.

APPENDIX : PROOFS OF THEOREMS

A.1 Assumptions

We list below all the assumptions we use in the development of our proofs. Note that the bandwidths h_1 and h_2 are by definition positive real numbers.

(A0) As $n \to \infty$, $h_1 \to 0$ and $nh_1 \to \infty$;

(A1) The kernel $K(.)$ is a bounded symmetric and continuous density function with finite 7^{th} moment ;

(A2) $f_X(x) > 0$ and $\ddot{f}_X(.)$ is bounded in a neighborhood of x where $f_X(.)$ denotes the marginal density function of X ;

(A3) $\ddot{m}(.)$ exists and is continuous in a neighborhood of x ;

(A4) $\sigma^2(.)$ has a bounded third derivative in a neighborhood of x and $\ddot{m}(x) \neq 0$;

(B0) As $n \to \infty$, $h_i \to 0$ and $\liminf nh_i^4 > 0$ for $i = 1, 2$;

(B1) The kernel $K(.)$ is a symmetric density function with a bounded support in $\mathbb{R}$. Further, $|K(x_1) - K(x_2)| \leq c|x_1 - x_2|$ for $x_1, x_2 \in \mathbb{R}$;

(B2) The marginal density function $f_X(.)$ satisfies $f_X(x) > 0$ and $|f_X(x_1) - f_X(x_2)| \leq c|x_1 - x_2|$ for $x_1, x_2 \in \mathbb{R}$;

(B3) $\mathbb{E}(Y^4) < \infty$;

(B4) $\sigma^2(x) > 0$ and the function $\mathbb{E}(Y^k|X = .)$ is continuous at x for $k = 3, 4$. Further, $\ddot{m}(.)$ and $\ddot{\sigma}^2(.)$ are uniformly continuous on an open set containing the point x ;

(C0) $f_X(.)$ has compact support $[a, b]$

Assumptions (A0) and (B0) are standard ones in kernel estimation theory. Some classical considerations on MSE or MISE (Mean Integrated Squared Error) lead to theoretical optimal constant bandwidths of order $n^{-1/5}$.

Assumptions (A1) and (B1) are directly satisfied by commonly used kernel functions. We can note that they require a kernel with bounded support, but this is only a technical assumption for brevity of proofs. For example, the Gaussian kernel can be used.

The assumption $f_X(x) > 0$ in (A2) and (B2) simply ensures that the experimental design is rich enough. The fact that (A2) also requires $\ddot{f}_X(.)$ to be bounded in a neighborhood of x is natural. The Lipschitz condition on f in (B2) is directly satisfied if f is sufficiently regular and with compact support.

Assumptions (A3), (A4), (B3) and (B4) are natural and ensure sufficient regularity to the conditional moments.

Assumption (C0) is made to make the presentation easier. It can be relaxed by means of the conventional truncation techniques used in real cases (Mack & Silverman (1982)). Nevertheless in practice, the input factors considered in sensitivity analysis are bounded and so have densities with compact support.

A.2 Proof of Theorem 1

This theorem is a direct consequence of the asymptotic behavior of the bias and variance in local linear regression.

Under assumptions (A0)-(A4), Fan et al. (1996) established that for a given kernel $K(.)$

$$\mathbb{E}_{\mathbb{X}}(\hat{m}(x)) = m(x) + \frac{1}{2}\mu_2 \ddot{m}(x) h_1^2 + o_P(h_1^2) \tag{6}$$

and

$$\mathrm{Var}_{\mathbb{X}}(\hat{m}(x)) = \frac{\nu_0 \sigma^2(x)}{f_X(x) n h_1} + o_P(h_1^2) \tag{7}$$

where $\mu_k = \int u^k K(u)du$ and $\nu_k = \int u^k K^2(u)du$. Now as the estimator $\hat{T}_1$ is

$$\hat{T}_1 = \frac{1}{n' - 1} \sum_{j=1}^{n'} \left(\hat{m}(\tilde{X}_j) - \hat{\bar{m}} \right)^2$$

we can write

$$\hat{T}_1 = \frac{1}{n' - 1} \sum_{j=1}^{n'} (Z_j - \bar{Z})^2$$

where $(Z_j)_{j=1,\ldots,n'} := (\hat{m}(\tilde{X}_j))_{j=1,\ldots,n'}$ and $\bar{Z} = \frac{1}{n'} \sum_{j=1}^{n'} Z_j$. By conditioning on the predictors $\mathbb{X}$, the sample $(Z_j|\mathbb{X})_{j=1,\ldots,n'}$ is an i.i.d. sample distributed as $Z_1|\mathbb{X}$ and the conditional bias of $\hat{T}_1$ can then be obtained through the classical formula for the empirical estimator of the variance :

$$\mathbb{E}_{\mathbb{X}}(\hat{T}_1) = \mathrm{Var}_{\mathbb{X}}(Z_1) = \mathbb{E}_{\mathbb{X}}(Z_1^2) - \mathbb{E}_{\mathbb{X}}(Z_1)^2.$$

Note that we can also compute its variance

$$\mathrm{Var}_{\mathbb{X}}(\hat{T}_1) = \frac{1}{n'} \left(\mathbb{E}_{\mathbb{X}}((Z_1 - \mathbb{E}_{\mathbb{X}}(Z_1))^4) - \frac{n' - 3}{n' - 1} \left(\mathrm{Var}_{\mathbb{X}}(Z_1) \right)^2 \right)$$

even though we do not use this result here (see Remark 1.).

As $\tilde{X}$ is independent of X and Y, we write

$$\begin{aligned}\mathbb{E}_{\mathbb{X}}(Z_1^2) &= \int \mathbb{E}_{\mathbb{X}}(\hat{m}(x)^2) f_{\tilde{X}}(x)dx \\ &= \int \left(\mathrm{Var}_{\mathbb{X}}(\hat{m}(x)) + \mathbb{E}_{\mathbb{X}}(\hat{m}(x))^2\right) f_X(x)dx.\end{aligned}$$

Considering assumptions (A3), (A4) and (C0) we then get using (6) and (7), in a similar way as for the standard MISE evaluation,

$$\begin{aligned}\mathbb{E}_{\mathbb{X}}(Z_1^2) &= \int m(x)^2 f_X(x)dx + \frac{\nu_0}{nh_1}\int \sigma^2(x)dx \\ &\quad +\mu_2 h_1^2 \int m(x)\ddot{m}(x) f_X(x)dx + o_P(h_1^2)\end{aligned}$$

and by the same arguments we also have

$$\mathbb{E}_{\mathbb{X}}(Z_1) = \int m(x) f_X(x)dx + \frac{1}{2}\mu_2 h_1^2 \int \ddot{m}(x) f_X(x)dx + o_P(h_1^2),$$

which finally leads to

$$\begin{aligned}\mathbb{E}_{\mathbb{X}}(\hat{T}_1) &= \mathbb{E}_{\mathbb{X}}(Z_1^2) - \mathbb{E}_{\mathbb{X}}(Z_1)^2 \\ &= \mathrm{Var}(\mathbb{E}(Y|X)) \\ &\quad +\mu_2 h_1^2 \left[\int m(x)\ddot{m}(x) f_X(x)dx \right. \\ &\quad \left. -\left(\int m(x) f_X(x)dx\right)\left(\int \ddot{m}(x) f_X(x)dx\right)\right] \\ &\quad +\frac{\nu_0}{nh_1}\int \sigma^2(x)dx + o_P(h_1^2) \\ &= \mathrm{Var}(\mathbb{E}(Y|X)) + M_1 h_1^2 + \frac{M_2}{nh_1} + o_P(h_1^2)\end{aligned}$$

where

$$\begin{aligned}M_1 &= \mu_2 \left[\int m(x)\ddot{m}(x) f_X(x)dx \right. \\ &\quad \left. -\left(\int m(x) f_X(x)dx\right)\left(\int \ddot{m}(x) f_X(x)dx\right)\right]\end{aligned}$$

and

$$M_2 = \nu_0 \int \sigma^2(x)dx.$$

A.3 Proof of Theorem 2

Similarly we first recall asymptotic results for the residual-based estimator of the conditional variance.

Under assumptions (B0)-(B4) Fan & Yao (1998) showed that

$$\mathbb{E}_{\mathbb{X}}(\hat{\sigma}^2(x)) = \sigma^2(x) + \frac{1}{2}\mu_2 \ddot{\sigma}^2(x) h_2^2 + o_P(h_1^2 + h_2^2)$$

and

$$\mathrm{Var}_{\mathbb{X}}(\hat{\sigma}^2(x)) = \frac{\nu_0 \sigma^4(x)\lambda^2(x)}{f_X(x) nh_2} + o_P\left(\frac{1}{\sqrt{nh_2}}\right)$$

where $\lambda^2(x) = \mathbb{E}((\epsilon^2 - 1)^2 | X = x)$ and μ_2 and ν_0 are as defined above. The estimator $\hat{T}_2$ can be written as

$$\hat{T}_2 = \frac{1}{n'}\sum_{j=1}^{n'} U_j$$

where $(U_j)_{j=1,\ldots,n'} := (\hat{\sigma}^2(\tilde{X}_j))_{j=1,\ldots,n'}$. As in the proof of Theorem 1, we then get the conditional bias and variance of $\hat{T}_2$:

$$\mathbb{E}_{\mathbb{X}}(\hat{T}_2) = \mathbb{E}_{\mathbb{X}}(U_1)$$

and

$$\mathrm{Var}_{\mathbb{X}}(\hat{T}_2) = \frac{1}{n'}\mathrm{Var}_{\mathbb{X}}(U_1).$$

As $\tilde{X}$ is independent of X and Y, we have

$$\mathbb{E}_{\mathbb{X}}(U_1) = \int \mathbb{E}_{\mathbb{X}}(\hat{\sigma}^2(x)) f_{\tilde{X}}(x)dx.$$

Considering assumptions (B4) and (C0) as in the proof of Theorem 1 we then get

$$\begin{aligned}\mathbb{E}_{\mathbb{X}}(\hat{T}_2) &= \mathbb{E}(\mathrm{Var}(Y|X)) + \frac{1}{2}\mu_2 h_2^2 \int \ddot{\sigma}^2(x) f_X(x)dx \\ &\quad +o_P(h_1^2 + h_2^2) \\ &= \mathbb{E}(\mathrm{Var}(Y|X)) + V_1 h_2^2 + o_P(h_1^2 + h_2^2)\end{aligned}$$

where

$$V_1 = \frac{1}{2}\mu_2 \int \ddot{\sigma}^2(x) f_X(x)dx$$

and using the same arguments

$$\begin{aligned}\mathrm{Var}_{\mathbb{X}}(\hat{T}_2) &= \frac{1}{n'}\left\{\mathbb{E}(\mathrm{Var}(Y|X)^2)\right. \\ &\quad +\mu_2 h_2^2 \int \sigma^2(x)\ddot{\sigma}^2(x) f_X(x)dx \\ &\quad -\mu_2 h_1^2 \left(\int \ddot{\sigma}^2(x) f_X(x)dx\right)\left(\int \sigma^2(x) f_X(x)dx\right) \\ &\quad +\frac{\nu_0}{nh_2}\int \sigma^4(x)\lambda^2(x)dx \\ &\quad \left.+o_P\left(h_1^2 + h_2^2 + \frac{1}{\sqrt{nh_2}}\right)\right\} \\ &= \frac{1}{n'}\left\{\mathbb{E}(\mathrm{Var}(Y|X)^2) + V_2 h_2^2 + V_3 h_1^2 + \frac{V_4}{nh_2}\right. \\ &\quad \left.+o_P\left(h_1^2 + h_2^2 + \frac{1}{\sqrt{nh_2}}\right)\right\}\end{aligned}$$

where

$$V_2 = \mu_2 \int \sigma^2(x)\ddot{\sigma}^2(x) f_X(x)dx,$$

$$V_3 = -\mu_2 \left(\int \ddot{\sigma}^2(x) f_X(x)dx\right)\left(\int \sigma^2(x) f_X(x)dx\right),$$

$$V_4 = \nu_0 \int \sigma^4(x)\lambda^2(x)dx.$$

REFERENCES

Cukier, R., Fortuin, C., Shuler, K., Petschek, A. & Schaibly, J. (1973), 'Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. I Theory', *The Journal of Chemical Physics* **59**, 3873–3878.

Efron, B. & Tibshirani, R. J. (1994), *An Introduction to the Bootstrap*, New York: Chapman and Hall/CRC.

Fan, J. & Gijbels, I. (1996), *Local Polynomial Modelling and its Applications*, London: Chapman and Hall.

Fan, J., Gijbels, I., Hu, T.-C. & Huang, L.-S. (1996), 'An asymptotic study of variable bandwidth selection for local polynomial regression', *Statistica Sinica* **6**, 113–127.

Fan, J. & Yao, Q. (1998), 'Efficient estimation of conditional variance functions in stochastic regression', *Biometrika* **85**, 645–660.

Jacques, J., Lavergne, C. & Devictor, N. (2004), Sensitivity analysis in presence of model uncertainty and correlated inputs, *in* 'Proceedings of SAMO2004'.

Jones, M., Marron, J. & Sheather, S. (1996), 'A brief survey of bandwidth selection for density estimation', *Journal of the American Statistical Association* **91**, 401–407.

Kennedy, M. & O'Hagan, A. (2001), 'Bayesian calibration of computer models (with discussion)', *Journal of the Royal Statistical Society Series B* **63**, 425–464.

Mack, Y. & Silverman, B. (1982), 'Weak and strong uniform consistency of kernel regression estimates', *Z. Wahrsch. Verw.Gebiete* **61**, 405–415.

McKay, M. (1996), 'Variance-based methods for assessing uncertainty importance in NUREG-1150 analyses', *LA-UR-96-2695* pp. 1–27.

Oakley, J. & O'Hagan, A. (2004), 'Probabilistic sensitivity analysis of complex models : a bayesian approach', *Journal of the Royal Statistical Society Series B* **66**, 751–769.

Owen, A. B. (2005), Multidimensional variation for quasi-monte carlo, *in* 'Fan, J. and Li, G., editors, International Conference on Statistics in honour of Professor Kai-Tai Fang's 65th birthday.'.

Ratto, M., Tarantola, S. & Saltelli, A. (2001), Estimation of importance indicators for correlated inputs, *in* 'Proceedings of ESREL2001'.

Ruppert, D. (1997), 'Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation', *Journal of the American Statistical Association* **92**, 1049–1062.

Ruppert, D., Wand, M. & Carroll, R. (2003), *Semiparametric Regression*, Cambridge: Cambridge University Press.

Ruppert, D., Wand, M., Holst, U. & Hˆssjer, O. (1997), 'Local polynomial variance function estimation', *Technometrics* **39**, 262–273.

Saltelli, A., Chan, K. & Scott, E. M. (2000), *Sensitivity Analysis*, Chichester: Wiley Series in Probability and Statistics.

Saltelli, A., Tarantola, S., Campolongo, F. & Ratto, M. (2004), *Sensitivity Analysis in Practice*, Chichester: Wiley.

Santner, T. J., Williams, B. J. & Notz, W. I. (2003), *The Design and Analysis of Computer Experiments*, New York: Springer Verlag.

Sobol', I. (1993), 'Sensitivity estimates for nonlinear mathematical models', *MMCE* **1**, 407–414.

Vilar-Fernández, J. & Francisco-Fernández, M. (2002), 'Local polynomial regression smoothers with AR-error structure', *TEST* **11**, 439–464.

Wand, M. & Jones, M. (1994), *Kernel Smoothing*, London: Chapman and Hall.